WCIT AI4C (AI for Charities) Group: Briefing Note #1

Generative AI & Large Language Models (LLMs):

Opportunities, Risks & Myths

PUBLISHED 1st AUGUST 2023*

Document Purpose & Intended Audience

This briefing note is intended to inform City Livery Companies and their associated schools and charities they help fund and support, about how to best understand and use generative AI tools such as ChatGPT.

The paper draws heavily on the work of WCIT's AI4C group, that supports over 40 charities on the deployment of AI technology via subject matter experts from the WCIT membership and its broader network.

To comment or to seek further information on AI or the work of AI4C, please contact: Paul Excell, Chair, WCIT A14C Group: paul.excell@member.wcit.org.uk

**Given the size and speed of the topic, this note is indicative only and will be updated on a regular basis.*

Synopsis

ChatGPT is an Artificial Intelligence tool that uses large data sets and high-performance computing to deliver believable responses to prompts that can augment many tasks, but as it is designed for plausibility over accuracy, the results should be used with caution and rarely used as a statement of truth.

What is ChatGPT?

ChatGPT is an AI-powered chatbot (https://en.wikipedia.org/wiki/Chatbot), that can convincingly write almost anything based on a limited brief or prompt, be it a question or command. It is a type of generative AI called a large language model or LLM which in turn is an artificial neural network (https://en.wikipedia.org/ wiki/Artificial_neural_network)

The recent success of Generative AI and LLMs to generate convincing text and images is down to three things – a vast amount of data, algorithms capable of learning from that data, and the computational power to do so.

GPT itself stands for 'Generative Pre-trained Transformer', which means it's a tool that can generate responses based on what it's already learned. It is a paid-for tool although there is a free version, that has some limitations.

This kind of AI has been in development for 20 years, after a software technique called 'deep learning' became popular (https://en.wikipedia.org/wiki/Deep_learning) combining vast datasets, powerful computers, running neural networks on graphics processing units or GPUs (https://en. wikipedia.org/wiki/Graphics_processing_unit) to recognise images, process audio and play games.

However, despite the responses feeling as if they were written by a human, they are not derived from any sentience or consciousness - in reality, this AI is a giant exercise in applied statistics, and should be understood and used as such.

ChatGPT is not the only game in town…

In this paper, we use ChatGPT, a service developed by OpenAI (partly owned by Microsoft amongst other investment firms), as a proxy for all Generative AI tools and services. For example, Google has a similar service to ChatGPT called Bard and there are many other 'generative AI' tools that use AI to create text, audio, video or images.

It should be noted that ChatGPT and other tools are commercial products. 'Free' access to these products is used either to tempt eventual purchase or to use the inputs provided to improve the product and perhaps sell on the data to other organisations who wish to use this information to better target their own commercial services.

Where does ChatGPT get its information?

ChatGPT uses a collection of Large Language Models (LLMs) which are numbered according to how advanced they are, with the free web version currently using GPT3.5 (www.chatgptui.com). LLMs are based on all sorts of sources including the web, books, social media and more. The resulting language dataset comprises hundreds of billions of words. The free version of ChatGPT is based on data collection that finished sometime last year, so it does not 'know' anything about the world after that time.

Can I trust the information I get from ChatGPT or Google Bard?

In a word, “no”.

But in the same way as you might use different sources to kickstart a research project or better understand what people are saying about a topic, ChatGPT and similar tools can be used in a way that can help get you started and find information that you weren’t aware of. The main caution is not to use a chatbot as your primary source for information, but instead take the answers it gives you and pursue them until you have found the real facts.

Remember it is for when you are stuck, not for when you are being lazy.

How does ChatGPT get its answers?

It’s important to understand how ChatGPT comes up with its answers. Crudely put, ChatGPT is very good at placing one word after another. It can do this because it has 'learned' so much from the massive data gathering exercises that form the basis of the model that powers it. As such, it does not 'know' anything at all; all it can do is put words that make sense one after another.

It’s often accurate but equally it can write utter nonsense. Its responses have been dubbed by many as ‘confidently wrong’ because the tone ChatGPT uses does not leave any room for doubt, even when it’s talking rubbish. These tools also have the tendency to 'hallucinate', where facts go completely out of the window and it states facts that are patently false, such as the fact that the current year is 2022 or that it loves the user.

It does not know how to communicate a level of confidence in what it has written, and if you attempt to probe into how it knows what it’s told you – for instance by asking for a list of citations – it will simply produce a list of things that look like citations but may not actually be real references at all.

Is all the information on ChatGPT verified?

The data used to train ChatGPT includes social media platforms, such as Reddit and Twitter – some of this content is going to be false, misleading and even harmful. Since some of what ChatGPT produces is based on what is on these sites and other similar platforms, this should give you some idea of what level of trust you should give it.

Other GPT-powered tools have different levels of credibility. For example, Bing’s chat tool provides citations and links to the sources for the facts it has presented. That’s not to say the sources themselves are accurate and there is a potential for citations to be ‘hallucinated’, as has been demonstrated by testing with Bing's system.

What does ChatGPT do with the information I enter into it?

Ask ChatGPT this question and it tells you that no information about what you enter into it is stored. However, the UK’s National Cyber Security Council (NCSC) points out that you should not enter sensitive information (such as personal details or company intellectual property) into chatbots, and not to perform queries that could be problematic if made public (for example sharing your secrets and asking ChatGPT to solve a personal dilemma).

Most terms and conditions on generative AI tools state they own all the IP (Intellectual Property) even if you have used it to generate copy from your blog or images from your photos. And, as generative AI tools become more prevalent and are used by companies for specific customer service purposes, the data you enter could be stored under the T&Cs of the companies you’re communicating with. Given that GPT-powered chatbots have the ability to misunderstand, there is always a chance therefore that the data held about you by these companies is incorrect, which could be a breach of the General Data Protection Regulations or GDPR.

How and where are ChatGPT and other generative AI being used?

Anyone can use ChatGPT for themselves – visit the website, sign up and start experimenting. Just bear in mind the limitations given above on how it works and can be used. ChatGPT is a language model, so it can’t generate art or images like some AI engines.

Various companies are experimenting with how to use ChatGPT in their services – for example, giving more personalised recommendations on retail websites. However, the bulk of usage is behind the scenes, often using the tool to process vast amounts of data to improve efficiency, or to combat fraud for example.

On the negative side, there are various ways in which this technology can be used to create convincing content that is either fake, misleading or even used for scams. It is early days, but evidence has already been uncovered of huge numbers of ‘content farms’ using entirely AI-generated content to lure people to the site and gain as much revenue as

possible from advertising. While human-based content farms have existed for some time, it is now possible to produce convincing text on an enormous scale at very low cost.

Generative AI could transform humans' relationship with computers, knowledge and perhaps even, themselves. They could solve big problems by developing new drugs, design new materials, or unlock new forms of energy. And on a more prosaic level, it can augment human capabilities, such as writing code and summarising documents, as below.

How should I use ChatGPT right now?

A smart aid to writing

ChatGPT has been compared to a keen, average-quality new graduate appearing at your organisation. They can do much of the grunt work, like poring over reports, while you do the things that only humans can do, for example relating to others.

AI is limited in this way at the moment because the underlying technology behind it, LLMs, are really only pattern-matching tools – they look at the words in an incomplete sentence and try to guess the next ones in the sequence. The results are an imitation of the texts they were trained on, rearrangements of word sequences that obey the rules of grammar; but as the LLM is reconstructing material that is slightly different to what already exists, it gives the impression of comprehension.

However, given this, they can be very useful in several ways to augment and speed up your work – for example it can scour through source material at high speed to find the perfect quote for a report, or to draft bullet points for a presentation, but as always with AI output, the facts need to be checked independently, as it is wholly reliant on interpreting information available on the internet.

It is especially useful when writing or thinking of a title for an email, as the title and the first paragraph are needed to grab the reader – ChatGPT can recommend headlines if given a few paragraphs, but much of what it provides can be clichéd or poor. It can therefore give clues but should not be used to provide the actual answer. As stated before, it is for when you are stuck, not lazy.

Checking computer code or looking-up spreadsheet shortcuts

ChatGPT is useful for computer coders to check their work and for spreadsheet users to find the right formula. However, this can be risky as proprietary code or actual corporate data entered into the LLM will potentially appear in someone else's responses. Some large organisations have therefore either banned or restricted ChatGPT for their software programmers and business analysts.

Everyday advice, but provide the context…

If you want to put up a shelf for example, ChatGPT can be a friendlier, cheaper alternative to tradespeople. However, to avoid overly general advice, the prompt used needs to be detailed, including what type of tools you have available, the kind of wall you want to put the shelf on, and so on, so that the LLM can have more context and provide better, more appropriate responses.

An educational assistant, with caveats

ChatGPT can be used as a research tool or as first draft essay writer and may even help with explaining maths homework.

However, any student or teacher or parent, should be aware of 'hallucinations' sometimes called 'confabulations', where AI sometimes makes stuff up with apparent confidence when asked a question it doesn't know the answer to. This is down to the fact that current versions of AI are designed to produce plausible not accurate results.

The Quality Assurance Agency for Higher Education has said it's too late to put 'the genie back in the bottle', and that universities (and schools) should promote 'positive' use of AI, rather than banning it. This means allowing the judicious deployment of the tech, as one might, say, a calculator or internet research, because it will be used when those students graduate into the world of work.

Tools exist claiming to identify whether work is AI-generated or not, but they're largely ineffective and biased against non-native speakers – this matters because 1.6 million schoolchildren in England have English as an additional language.

Boosting admin work

AI is a unique technology with its broad applicability – it can be used in every walk of life and every industry, and every function in a company. Indeed, if you're an office worker with highly specialised knowledge, AI can help you do the tasks you don't want to, like crafting the content of an email, as long as the response is used as a guide, not the solution.

What should I do to make ChatGPT safer for me?

All use of technology services come with health warnings and below are listed some obvious ones for AI:

Check the Privacy Policy

Most of us just click 'accept' when presented with a Privacy Policy as we have little time or interest in wading through 'legalese'. However, it is worth reading the Privacy Policy before you use any Generative AI service.

By default, anything you talk to ChatGPT about could be used to help its underlying large language model "to learn about language and how to understand and respond to it". Personal information may be used to improve OpenAI's services and to develop new programmes and services – in short, it has access to everything you do on their services, and you are trusting the Company not to do anything dodgy with it. It is a similar story with Google's privacy policy. However, as with any data that Google gets from you, Bard data might be used to personalise the ads you see.

Watch what you share

With this in mind, you need to be aware that anything you put into an AI tool is likely to be used to refine the AI and then used as the developer sees fit. You should therefore be circumspect about what you enter into these engines. This is particularly true of image-based AI tools – you should only enter images of yourself for example if you are happy to see AI-generated versions of your face show up in others' creations.

In terms of text, DO NOT put in any personal, private or sensitive information. No matter how tempting it can be to summarise a document or write a letter with your address on it, this is not something that you want embedded into the training data. This is of course something that the companies caution against.

In short, if you don't want something to appear in public or be used in an AI output, keep it to yourself.

Change the settings

If you have decided you are ok with the privacy policy and you are not over-sharing, you need to explore and adjust the privacy and security settings in your AI tool of choice, which are relatively easy to access and change. Google for example allows you to have all data automatically erased after a user-defined set period of time, to manually delete it, or allow them to keep it forever.

What is the legal status of data from Generative AI?

At the time of writing (July 2023), there are no specific regulations either in the UK or worldwide for generative AI tools like ChatGPT but some jurisdictions (e.g., EU AI Act) have legislation pending. There has been a recent voluntary Code of Conduct agreed by the US Government and some leading AI companies (see below) but the US Government has reserved the right to propose future binding regulations at some point. However, other laws could well be applied to the responses LLMs produce, for example:

Copyright & Defamation

As ChatGPT generates text that is largely similar to a copyrighted source, it could be in breach of copyright law. In addition, there are already examples of ChatGPT defaming individuals, for example, stating they were involved in crimes they did not commit, which could result in legal action.

Data Protection

AI, like ChatGPT, has to obey data protection rules on the use and storage of personal data. Indeed, Italy's data protection regulator was quick off the mark to ban ChatGPT under data protection grounds to prevent personal data being stored and then shared back to other users. The ban was subsequently lifted after OpenAI "addressed or clarified" the issues that the regulator had raised.

The White House AI Watermarking Code of Conduct

On July 21st seven companies developing AI tools have made a commitment to 'watermarking' content created by their systems (https://en.wikipedia.org/wiki/Digital_watermarking) to combat the spread of disinformation and 'deep fakes'. Amazon, Anthropic, Google, Inflection AI, Meta, Microsoft and OpenAI have struck a deal with President Biden's administration to submit their technology to external testing, share information about their systems and to invest in combating cybercrime. The White House confirmed that it intends to introduce laws to combat the threats posed by the technology but said that the commitments make a critical step forward in developing responsible AI. It is notable that several companies developing AI image generation tools were not part of this announcement many of whom have submitted their software to open source.

APPENDIX: How does Generative AI work?

While very complex, at a high level, generative AI is trained using a set of data, usually taken from the internet. The AI ingests that data — the LLM that the free version of ChatGPT uses was trained on about as much data as a mid-range laptop contains — and begins sorting through it to try to understand it, building patterns in language.

This can be as simple as learning that "I love you" is a more common phrase than "I love doing my taxes", meaning that if you ask it to complete the phrase "I love", it's more likely to choose the former than the latter. But it also makes connections between subjects, meaning that if you ask ChatGPT to write a script for East Enders in the style of Shakespeare, it could probably do a decent job.

To understand how it works on a slightly more detailed level, it is important firstly to understand that neural networks, the computing technique used in LLMs, can only 'read' numbers, so words have to be converted into numbers. As such, when you make a query or prompt into ChatGPT, the words are converted into numbers.

Secondly, any text that commonly occurs together or has similar meaning, is further grouped together in what is known as 'tokens' – this is both whole words like 'love' and 'are', affixes like 'dis' or '-ised' or punctuation like '?'.

LLM models can handle very large token inputs – the more text the model can take in, the more context it sees, and the better its answers will be. The tokens are then assigned 'definitions' by embedding them into a 'meaning space' where words with similar meanings are located in nearby or adjacent areas.

Thirdly, the LLM deploys its 'attention network' to make connections between different parts of the prompt. The model learns these associations from scratch during billions of 'training runs'. It is vital to note that LLMs only understand language in a purely statistical, rather than grammatical, way – more abacus than mind.

Fourthly, once the prompt has been processed, the LLM initiates a response using its attention network to attach a probability that the token is the most appropriate one to use next in the sentence it is generating. However, the token with the highest score is not always the one chosen, based on how 'creative' the model has been told to be by its operators.

Finally, the LLM generates a word and feeds the result back into itself with the first word generated by the prompt alone. The second word is generated by including the first word in the response, the third by including the two previous and so on, in a process called 'autoregression', which repeats until the LLM has finished.

This all happens very fast, in a way that can make someone believe that the response is 'intelligent', whereas the correct understanding is that the response is a very smart exercise in 'applied statistics'.

In short, LLMs are just representations of the distributions of words in texts that can be used to product more words. Given they have no experience of real life or human communication they offer nothing more than the ability to parrot things they have heard in training. This is nothing like thought or intelligence as we know it in a human sense.

The way that LLMs work accounts for the 'hallucinations' (sometimes called 'confabulations') that have been seen where the LLM has derived facts from data that are untrue or do not exist. Some of these are nonsense such as the incorporation of fictional deeds in biographical sketches.

LLMs are designed for coherence rather than truth. This very plausibility aligned with a lack of veracity shows that to overly rely on the output of LLMs is naïve.

SOURCES

The information contained in this briefing note was derived from the following publications and sources which are wholly acknowledged and thanked:

The Economist; Which? Magazine; BCS Magazine; The Financial Times; BBC Online; The Times; The New York Times; and AI4C and WCIT expert members

The document was produced without interaction with ChatGPT or any other LLM.

DISCLAIMER